

AUTOMATED SYNTHESIS OF PREDICTION MODELS FOR NEURAL NETWORK BASED MYOCARDIAL INFARCTION CLASSIFIERS

J. A. Lopez, C. D. Nugent, N. D. Black, A. E. Smith

Medical Informatics, Faculty of Informatics, University of Ulster at Jordanstown

Abstract- Parameter and architectural selection for Multiple Layered Perceptron (MLP) classifiers involve a number of heuristic design procedures. The aim in the design process of such classifiers is to achieve maximum generalization and avoid over-fitting of the training data. It has been the objective of this study to develop a symbolic prediction model to calculate the point at which training should cease for a given Neural Network (NN) based 12-lead ECG classifier to ensure maximum generalization. This prediction model has been obtained by means of Genetic Programming (GP), where a GP individual has been evolved to generate a symbolic model that predicts the optimal number of training epochs for three different ECG myocardial infarction classifiers: Anterior Myocardial Infarction (AMI), Inferior Myocardial Infarction (IMI), and Combined Myocardial Infarction (CMI). The GP model demonstrated to be a very accurate method showing no significant differences between the optimal number of epoch values and the predicted values for both: train and test data sets for the three aforementioned pathologies.

Keywords - Genetic Programming, Neural Networks, Myocardial Infarction

I. INTRODUCTION

The main objective when classifying the ECG is to allocate patients into a probable list of cardiac pathologies. This classification process can be described within three functional modules: beat detection, feature extraction/selection, and classification [1, 2]. A NN based ECG classifier consists of artificial neurons assembled together in successive layers; such a NN structure is referred to as an MLP. The number of nodes and the number of hidden layers in an MLP are not fixed and are highly application specific [3]. An ECG classifier based on an MLP must first undergo training through a process of supervising learning. Following training, as determined by the training algorithm, the network is exposed to a set of unseen data in order to evaluate the performance of the network. When employing an MLP as a classifier of an unknown ECG signal, the input to the network is the input feature vector as produced following the stages of beat detection and feature extraction. The number of nodes in the hidden layer and the number of hidden layers themselves are varied during different attempts of training. Each neuron in the output layer represents a specific diagnostic class. Therefore, based on the input feature vector presented to the network, the output neuron with the largest output value is indicative of the presence of a specific diagnostic class. The current work is related to a previously developed classification framework for 12-lead ECGs based on a bi-group NN configuration (BGNN)

[4]. In the case of the aforementioned architecture (BGNN) only one neuron is associated with the output layer, in other words one classifier is able to predict the presence or absence of a particular pathology. The training and selection of the network is a heuristic procedure and many efforts have been achieved to produce the optimal classifier. A well-designed MLP will show high levels of generalisation if a correct input-output mapping is obtained even when the input is slightly different from the examples used to train the network. Many issues have been associated with the design process of an MLP, but the problem of locating the point at which the network is considered to be trained is still regarded as unresolved. Conventional methods will cease training whenever the point at which the minimum error for the training data is reached. These methods involve many risks, as it is not possible to know when to stop training for maximum generalisation and avoid over-fitting. Over-fitting occurs when the NN memorises the training data, and subsequently if unseen data is presented poor generalisation is attained. For this reason, it is possible to over-fit a NN if the training of the network is not stopped at an optimal point.

II. METHODOLOGY

The database used in this study comprises six different parameters, one of them being the dependent variable. These parameters have been identified as the variable design parameters in the development of each of the BGNNs and for this reason these are the most likely variables to potentially effect the position at which the point of maximum validation performance occurs. The five independent parameters are used as the input to the prediction model, and these are as follows:

1. Number of nodes in the hidden layer (n).
2. Feature Selection method employed (fs)
3. Number of files in training set (N)
4. Size of input feature vector (s)
5. Number of epochs for the NN to attain maximum performance during training (m).

Fig.1, shows a block representation of the prediction model where the number of epochs at which the NN attains maximum performance (**number of epochs**) is represented as the output or the dependent variable of a non-linear symbolic model, as follows:

$$number_of_epochs = F(n, fs, N, s, m, a_1, \dots, a_n) \quad (1)$$

Report Documentation Page

Report Date 25OCT2001	Report Type N/A	Dates Covered (from... to) -
Title and Subtitle Automated Synthesis of Prediction Models for Neural Network Based Myocardial Infraction Classifiers		Contract Number
		Grant Number
		Program Element Number
Author(s)	Project Number	
	Task Number	
	Work Unit Number	
Performing Organization Name(s) and Address(es) Medical Informatics, Faculty of Informatics, University of Ulster at Jordanstown		Performing Organization Report Number
Sponsoring/Monitoring Agency Name(s) and Address(es) US Army Research, Development & Standardization Group (UK) PSC 802 Box 15 FPO AE 09499-1500		Sponsor/Monitor's Acronym(s)
		Sponsor/Monitor's Report Number(s)
Distribution/Availability Statement Approved for public release, distribution unlimited		
Supplementary Notes Papers from the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 25-28 October 2001, held in Istanbul, Turkey. See also ADM001351 for entire conference on cd-rom., The original document contains color images.		
Abstract		
Subject Terms		
Report Classification unclassified	Classification of this page unclassified	
Classification of Abstract unclassified	Limitation of Abstract UU	
Number of Pages 4		

In (1) F is a non-linear function represented by a symbolic expression with arithmetic functions of plus, minus, product, and protected division and (a_1, \dots, a_n) is a predefined vector of float type constants.

Only BGNN based performance data for myocardial infarction classification was used in this study for AMI, IMI, and CMI. Each data set was segmented with two thirds allocated to training data and one-third as test data for later evaluation of the GPs performance. In this work, GP has been investigated and applied for the development of a symbolic prediction model that matches equation (1) and models the black box representation illustrated in Fig 1.

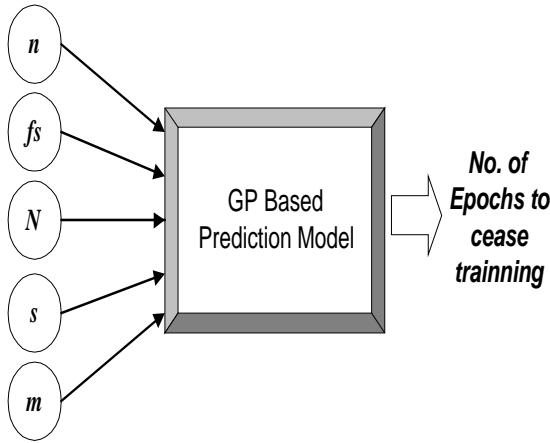


Fig 1. GP based Prediction model to determine number of epochs to cease training.

III. GENETIC PROGRAMMING FOR THE IMPLEMENTATION OF THE PREDICTION MODEL

In this section the proposed GP system for the prediction model already discussed is described. GP is an automatic method for creating a working computing program for a high level statement of a problem. GP can be defined as a search method based on natural selection rules [5, 6]. In GP a population of candidates to solution programs is evolved. An individual of the population (a program) is, the most of the time, represented as a tree where some nodes are functions and some others are terminal symbols. In order to obtain a good individual (the program that solves the problem), appropriated functions and terminal sets have to be chosen. A fitness function is used to evaluate the performance of each individual in the population. Following this, genetic operators such as crossover, reproduction and mutation are applied to each individual and then some of the fittest individuals are selected to survive in further generations. This process repeats iteratively until a good candidate solution is found or a predefined maximum number of generations are reached. A population of 3000 individuals was evolved with a function set consisting of arithmetic functions as follows: Addition (+), Subtraction (-), Protected division (/) and Product (*).

The function set can be denoted as:

$$\Gamma = \{+, -, *, /\} \quad (2)$$

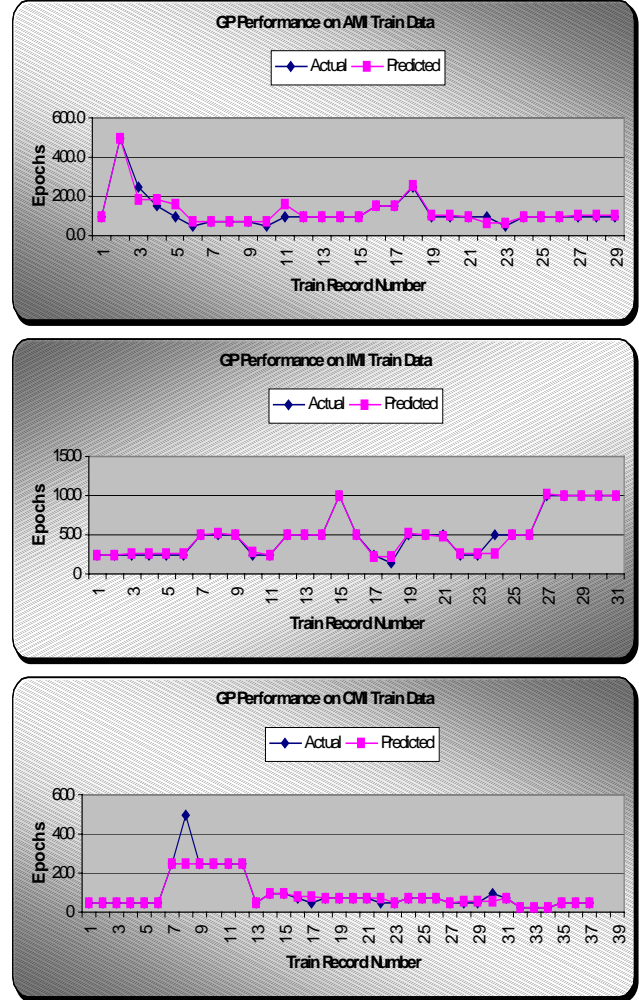


Fig 2. GP performance for the three myocardial infarction models (AMI, IMI and CMI) in data used for the train dataset.

The terminal set consisted of the following variables:

- A set of random float type constants between: 0.0 and 5.0 (a_1), 0.0 and 50.0 (a_2) and 0.0 and 500.0 (a_3).
- Variable n : Number of nodes in the hidden layer.
- Variable fs : Feature Selection method employed.
- Variable N : Number of files in training set.
- Variable s : Size of input feature vector.
- Variable m : Number of epochs for NN maximum performance during training.

Then the terminal set can be denoted as:

$$T = \{a_1, a_2, a_3, n, fs, N, s, m\} \quad (3)$$

The Fitness Function: The fitness function was based on absolute raw errors for the desired output parameter (number of epochs) and the complexity of each individual to avoid large individuals and ensure generalization.

IV. RESULTS

Following the evolution process three individuals were found for each of the aforementioned pathologies:

- For AMI an individual with raw fitness 340.5 and complexity 127.
- For IMI an individual with raw fitness 487.2 and complexity 127.
- For CMI an individual with raw fitness 401.0 and complexity 151.

Individuals were synthesized in the form of LISP type S-expressions. Comparison between desired and actual values of epochs for the three myocardial infarction models (AMI, IMI, and CMI) for both training and testing datasets are illustrated in Fig 2 and Fig 3. Performance was measured and statistically validated using the Wilcoxon's signed rank sum test for paired data. These results are presented in Table 1, showing no significant difference at the $p=0.05$ level for IMI test and CMI test. The AMI test result was just marginally significant (it is slightly overestimating the epochs) and this is reflected in the differences in the mean +ve and -ve ranks.

TABLE I
WILCOXON'S SIGNED RANK SUM RESULTS FOR GP PREDICTION MODELS

GP	No. of Cases	Mean Rank -ve	Mean Rank +ve	z-value	2-tailed sig
AMI Train	29	13.08	16.56	-1.027	$p=0.304$
AMI Test	15	4.17	10.56	-1.989	$p=0.047$
IMI Train	31	16.17	15.89	-1.058	$p=0.290$
IMI Test	15	9.67	6.89	-0.114	$p=0.910$
CMI Train	37	19.56	18.47	-0.008	$p=0.994$
CMI Test	14	9.38	5.00	-1.412	$p=0.158$

V. CONCLUSION

GP has demonstrated, in the current study, to be a very good method in the given NN reengineering problem. Fig 3 and Table 1 show that a GP based prediction model not only performs very well with training data, but also demonstrates high generalization capabilities. The result from this study shows that it is possible, given the design parameters of a NN ECG classifier, to predict the point at which training should cease for maximum generalization. This is a very powerful result as it indicates promise to alleviate the lengthy and uncertain design process of MLP classifiers.

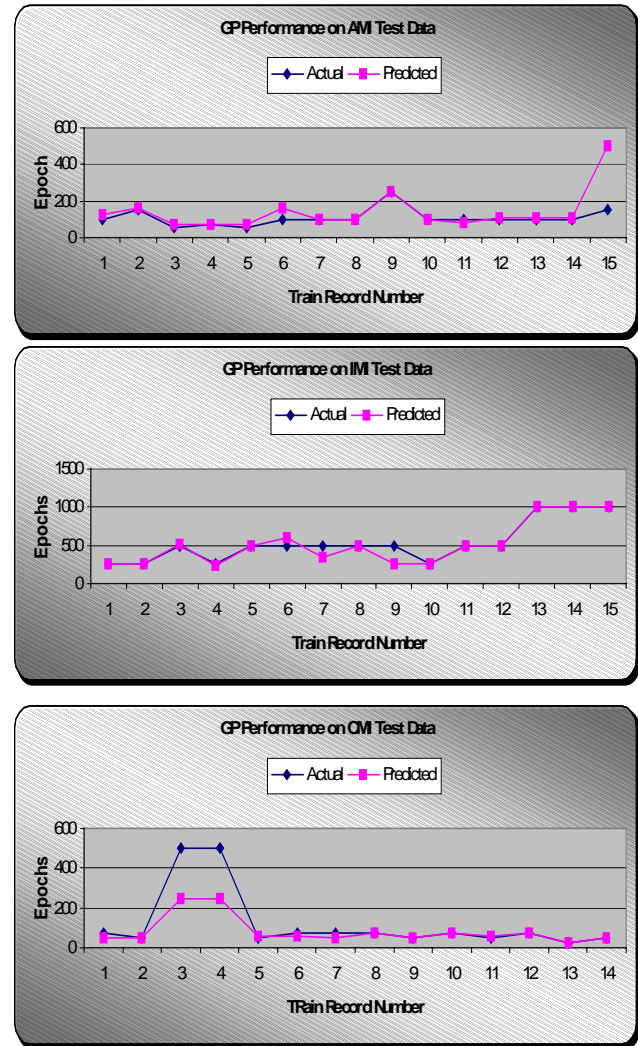


Fig 3. GP performance for the three myocardial infarction models (AMI, IMI and CMI) in data used for the test dataset

REFERENCES

- [1] C. D. Nugent, J. A. C. Webb, G. T. H. Wright and N. D. Black, "Electrocardiogram 1: pre-processing prior to classification". *Automedica*, vol. 16, pp. 263-282, 1998.

- [2] C. D. Nugent, J. A. C. Webb, N. D. Black and G. T. H. Wright. "Electrocardiogram 2: Classification". *Automedica*: vol. 17, pp: 281-306. 1999.
- [3] D. A. Zahner, and E. Micheli-Tzanakou, "Artificial neural networks: Definitions, methods, applications". In: *Ed. Bronzino, J.D., The Biomedical Engineering Handbook, CRC Press*. pp. 2699-2715. 1995.
- [4] C. D. Nugent, J. A. C. Webb, N. D. Black, G. T. H. Wright and M. McIntyre. "An Intelligent Framework for the Classification of the 12-Lead ECG". *Artificial Intelligence in Medicine*: vol. 16, pp: 205-222. 1999.
- [5] J. R. Koza, "Genetic Programming: On the Programming of Computers by Natural Selection". *MIT Press, Cambridge, MA*. 1992.
- [6] J. R. Koza, "Genetic Programming II: Automatic Discovery of Reusable Programs". *MIT Press, Cambridge, MA*. 1994.